共享住宿与酒店用户评论文本的跨平台比较研究：基于 LDA 的主题社会网络和情感分析*

■ 池毛毛[1,2] 潘美钰[1] 王伟军[3]

[1] 华中师范大学信息管理学院 武汉 430079 [2]华中师范大学湖北省电子商务研究中心 武汉 430079

[3] 华中师范大学青少年网络心理与行为教育部重点实验室 武汉 430079

摘 要：[目的/意义] 共享住宿与酒店预定平台可能同时存在替代性和互补性，但这种替代性和互补性分别体现在哪些产品和服务上当前文献还缺乏探讨，需要进一步开展跨平台的比较研究。[方法/过程] 选取携程酒店预定平台和小猪短租平台为实验对象，采集北京市相关房源的 86 635 条用户评论文本，结合 LDA 模型、主题社会网络和主题情感分析方法对用户文本评论进行跨平台比较分析。[结果/结论] 研究发现两大平台用户在评论主题、主题社会网络和主题情感上的异同之处，从微观用户评论角度解释了两大平台在产品和服务上的替代性和互补性。本文结果为平台管理者进行住宿产品和服务的开发和改进提供重要的实践借鉴。

关键词：跨平台比较 文本主题挖掘 社会网络分析 情感分析

分类号：TP391.1 F719.2

DOI：10.13266/j.issn.0252－3116.2021.02.011

共享经济模式近些年在国民经济各行各业得到了迅速渗透，特别是在旅游行业，共享住宿（例如 Airbnb、途家、小猪短租等）作为一种新的业态已经成为旅行者住宿选择的重要方式之一。共享住宿平台的发展对于酒店预定平台已经产生一定冲击，抢占了酒店的部分市场。根据《中国共享住宿发展报告 2019》，2018 年共享住宿市场占全国住宿业的 6.1%，未来三年共享住宿市场规模将会继续保持 50% 的增长速度。然而，不同于传统酒店业，共享住宿服务于全新的客户群，并会刺激传统住宿业（包括酒店业）转型升级。2015－2018 年我国共享住宿的发展对住宿业年均增长的拉动作用为 2.1 个百分点。因此，共享住宿行业和酒店行业的产品和服务可能同时存在替代性和互补性，而这种替代性和互补性分别体现在哪些产品和服务上，需要进一步开展跨平台的比较研究。

现有关于共享住宿和酒店预定平台之间相互关系的文献主要是从宏观经济角度分析新生业态——共享住宿对于酒店行业的影响。研究发现 Airbnb 的入驻对于当地传统酒店行业是一种冲击，酒店的销量和定价均受到影响[1-2]。另有相关文献基于酒店预定平台或共享住宿平台的用户文本评论数据，从微观用户评论角度挖掘平台用户关注的主题维度，例如硬件提供、服务质量、环境、饮食和性价比等酒店预定平台的主题[3]以及住宿设施、房东相处、位置、类家性和原真性等共享住宿平台的主题[4-5]。然而，当前文献主要基于宏观经济影响角度探究平台间的作用机制，微观用户评论角度文献则重点关注单一平台，尚缺乏对于产品和服务系统性的跨平台比较研究。

基于上述实践与理论动机，本文采集酒店预定平台（携程酒店平台）和共享住宿平台（小猪短租平台）在北京市的 86 635 条房客文本评论数据，并整合 LDA 主题模型、社会网络分析（SNA）和情感分析方法展开跨平台的用户评论文本主题分析。研究发现两个平台在用户评论主题、主题社会网络和主题情感倾向方面的异同点。本文结果为相关住宿平台管理者进行产品和服务的开发和改进提供重要的理论指导和实践借鉴。

* 本文系国家自然科学基金青年项目“电商平台演化对平台绩效的影响机理研究：基于复杂适应系统的视角”（项目编号：71801104）研究成果之一。

作者简介：池毛毛（ORCID：0000－0003－2726－5933），副教授，博士，硕士生导师，E-mail：chimaomao@aliyun.com；潘美钰（ORCID：0000－0001－7591－3726），本科生；王伟军（ORCID：0000－0003－4948－0634），教授，博士生导师。

收稿日期：2020－04－13 修回日期：2020－07－20 本文起止页码：107－116 本文责任编辑：杜杏叶

1　相关研究

1.1　酒店预定平台与共享住宿平台关系的相关研究

目前对于酒店预定平台和共享住宿平台相互关系的文献集中在从宏观经济影响角度分析共享住宿对于酒店业的作用机制，将共享住宿（例如 Airbnb）视为一种破坏性创新，研究其对于传统酒店定价和销量的影响，并提出共享住宿和传统酒店之间可能存在一定的互补性或替代性。相关文献发现 Airbnb 的入驻对当地酒店业绩效有负面影响，即共享住宿与酒店之间存在替代性。例如 K. L. Xie 等通过分析 Austin 城市的 86 家酒店数据，发现 Airbnb 房源的供应数量会对酒店绩效产生负向作用，并且指出该负向作用并不会受到酒店的质量属性（星级评定）的影响[2]。类似地，T. Dogru 等发现伦敦和巴黎等城市十年间（2008 – 2017 年），Airbnb 房源供应数量对酒店收入和入住率均产生了负向影响[6]。但是，I. Blal 等通过分析旧金山的 101 家酒店数据，发现 Airbnb 的入驻和酒店业绩效之间同时存在替代性和互补性，其中互补性主要体现在 Airbnb 房源的供应数量不会影响酒店收入，替代性体现在 Airbnb 的平均价格和用户满意度与酒店绩效呈负相关[1]。

为了进一步挖掘用户对于相关旅游预定平台的评价信息，相关文献从微观角度分别针对酒店预定平台和共享住宿平台上的用户评论进行文本分析。不同学者提炼出了酒店预定平台用户关注的主题维度。例如，赵学锋等在携程酒店平台上选取北京、上海等 150 家酒店的前 10 条评论，基于 DBSCSN 聚类方法，总结出酒店预定平台上用户关注的五个主要维度，包括硬件提供、服务质量、环境、饮食和性价比[3]。吴维芳等则采集了拉斯维加斯的酒店评论数据，基于 K-means 聚类方法进一步发现位置、网络提供和清洁度三个主题维度[7]。相关研究又发现酒店人员服务比其他特征维度对用户影响更大[8]。另有文献则试图挖掘共享住宿平台用户所关注的主题维度。例如，M. M. Cheng 等通过采集 Airbnb 平台上悉尼市的评论文本，利用词共现的分析方法发现用户主题主要包括住宿设施、房东相处和位置[4]。卢长宝等结合高频特征词的语言情境，将 Airbnb 平台上房客评论关注主题划分为八个维度，并提出类家性和原真性（文化氛围）两个主题[5]。

综上所述，当前围绕酒店预定平台与共享住宿平台二者关系的研究主要基于宏观经济影响（即共享住宿作为破坏性创新对于传统酒店业的影响）和微观用户评论（即平台用户评论主题的挖掘）两个方面，但是当前文献尚存在如下三点不足：首先，当前文献对于酒店预定和共享住宿平台的关系研究，主要是从宏观经济影响的角度分析共享住宿对传统酒店业的作用，缺乏从用户评论角度对于两个平台产品和服务系统性的微观比较研究；其次，现有基于旅游平台上文本评论的研究局限于单一平台，即分别针对酒店预定平台的携程酒店和共享住宿平台的 Airbnb 的用户文本评价进行主题挖掘，尚缺乏针对跨平台用户评论文本的比较分析；最后，当前相关旅游文献对文本评论的分析方法上主要基于主题挖掘分析，缺乏对评论主题关联关系与情感分析结果的细化分析，难以细粒度地呈现不同平台用户所关注主题的具体差异。

1.2　主题挖掘研究：基于 LDA 的主题社会网络和情感分析

主题挖掘方法能够有效地识别文本主题，挖掘用户在线观点。目前常用的主题挖掘有两类：一类是依赖于文本相似性的传统主题聚类模型[9-10]，该类模型应用广泛、操作简单，但模型聚类结果依赖于文本之间的距离[11]，通过词语频率和统计方法分析内容获取的结果较为杂乱，并不能真实表达评论者意图；另一类是概率主题模型，如 LDA 模型等[12-13]。概率主题模型能高效地挖掘出大量文档中所包含的主题信息，由于 LDA 模型对于文本长度没有严格的限制[14]，近几年被广泛应用于用户评论主题与关注热点的识别。例如，在微博类的热点挖掘方面，相关文献基于新浪微博的热门微博数据，利用 LDA 模型结果进行分类推断以及主题挖掘[15]，另有研究结合 LDA 模型对网络热点事件进行话题抽取，探究网络事件的演化分析[16]；在电商在线评论领域，有研究基于 LDA 模型挖掘出生鲜电商行业影响用户满意度的关键因素，如包装、新鲜度、物流运输等[17]；在旅游在线评论中，相关文献利用 LDA 模型对“庐山旅游”在线评论进行主题聚类，并针对不同需求类型提出产品/服务改进的建议[13]。综上可以看出 LDA 模型的应用广泛，并能有效且显著地发现短文本的主题特征。因此，本文将选用概率主题 LDA 模型识别酒店预定和共享住宿平台上文本评论的主题。

LDA 模型被称为三层贝叶斯概率模型，属于一种文档主题生成模型。用户评论文本也会包括一定概率的关注内容的主题集合，而该主题包括一定概率的词语集合，评论文本到主题、主题到词语均服从多项式分布，如公式(1)和公式(2)所示：

$$z \sim Multi(z|(\overline{\theta_m}))\qquad \text{公式(1)}$$

其中 z 代表的是第 m 条评论中产生的主题随机变量，$\overline{\theta_m}$代表的是在 M * K 的矩阵中第 m 条评论的多项式分布参数，M 代表评论数量，K 代表主题数量。

$$w \sim Multi(w \mid \overline{\theta_k}) \qquad 公式(2)$$

其中 w 代表的是第 k 个主题产生的词语随机变量，$\overline{\theta_k}$代表在 K * V 的矩阵中第 k 个主题的多项式分布参数，V 代表词语数量。虽然概率主题模型可以帮助识别文本的主题，但是单靠主题挖掘难以细粒度地识别主题间的关联关系以及主题的情感倾向。在主题挖掘的基础上，有文献利用主题词之间的共现关系，挖掘词与词之间的关系网络，并将基于 LDA 和 SNA 的文本挖掘模型应用至信息检索[18]和新闻热点识别[14]，且有效性均得到了验证。除了识别用户关注信息特征维度关系，用户情感倾向也有助于进一步理解用户对该主题的态度。因此，结合主题挖掘和情感分析将有助于识别出用户对事件的响应态度和关注重点。

综上，区别于以往"主题识别 – 共词分析"[19]和"主题识别 – 情感分析"[20]的处理文本内容的方法，本文在主题识别的基础上，同时引入社会网络分析方法（SNA）构建主题 – 主题和主题 – 特征词的社会网络，并运用情感分析方法量化用户对于主题的情感倾向。其中，SNA 为分析主题词重要程度与关联词语提供支持，词语的共现程度也能从侧面反映关联关系[21]。该方法能够识别平台用户的核心主题与主题间的社会网络关系，利于挖掘用户评论主题的关联关系。情感分析方法则是通过使用自然语言处理技术从文本数据中识别出用户主观的情感、观点和态度的过程[22]。情感分析方法主要包括基于机器学习的情感分类方法与基于情感词典的语义分析方法[23]。住宿平台的评论文本属于短文本，且通常评论句子表达不规范。由于机器学习对表达情感的情感符号相对不敏感，基于机器学习的情感分析需要依赖大量的训练样本和人工干预，效率相对较低[20]。因此，本文将基于情感词典的情感分析方法对跨平台的用户评论进行主题情感分析。具体来说，通过建立主题与情感态度之间的联系，本文重点挖掘用户文本评论各主题下的情感倾向，进而分析酒店预定与共享住宿平台的主题情感差异。

2 研究方法

2.1 数据来源

本研究的数据来源于酒店预定平台和共享住宿平台 2015 年 11 月 – 2019 年 11 月的在线评论，并选择该平台上北京市作为评论数据的采集对象。北京作为我国政治和文化中心，吸引着世界各地的旅客，在线用户评论丰富且全面，能够真实反映我国传统酒店和共享住宿行业的现状。酒店预定平台的评论数据爬取自携程酒店平台，共享住宿平台的评论数据爬取自小猪短租平台。其中，①携程酒店平台是国内在线酒店预定行业的行业标杆。携程一直占据着在线住宿预订市场商业价值第一的位置，且保持着强劲的竞争力。由于酒店行业的评论数目较多，本文爬取各市辖区首页的酒店及其相应的评论，最终得到有效数据 70 家酒店及 55 761 条房客文本评论。②小猪短租平台是国内在线短租共享住宿行业的新星，该平台凭借其人情化的品牌服务赢得了众多用户的青睐。目前该平台的全球房源数量超过 80 万，城市覆盖超过 710 座。本文在小猪平台爬取 5 534 间房源，由于部分房源无评论，最终有效数据 2 635 间房源及对应的 30 874 条房客文本评论。因此，本研究在以上平台中获取文本评论数据共计 86 635 条。

2.2 评论文本主题挖掘模型

本研究流程主要分为如下五个具体步骤。①在携程酒店平台和小猪短租平台采集评论数据，形成后文分析所需的跨平台评论文本库；②文本数据预处理，主要包括文本分词、去除停用词并对词性进行标注等；③基于 LDA 主题模型对文本评论进行聚类，挖掘评论主题；④基于不同主题间的关系与主题内部特征词间的关系构建社会网络，并对各个主题进行基于情感词典的情感分析；⑤对比酒店预定平台与共享住宿平台的 LDA 聚类、社会网络分析和情感分析结果，进行比较分析。具体流程如图 1 所示：

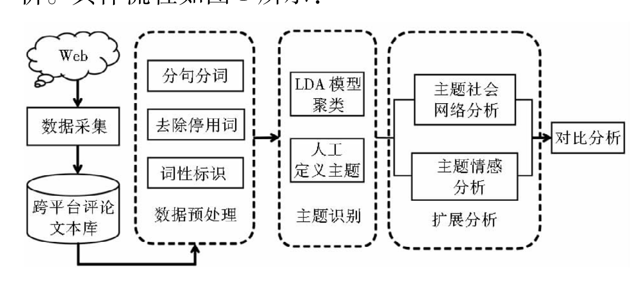


图 1　研究流程

本部分重点对在相关研究流程中采用的主要方法进行介绍。①数据预处理。本部分主要是对评论内容进行去分词。为了提高分词效率，保证分词的准确性和完整性，本文采用自动分词与人工处理相结合的方法来进行处理文本。停用词库的构建主要是选取哈工大停用词表（767）和百度停用词表（1 395）进行停用词去重整合。最后利用 Python 中 Jieba 包完成文本信息

的分词过程。②LDA 主题建模。文本的分类主题建立在 LDA 模型的聚类结果,由于 LDA 主题抽取的效果和主题数确定有直接的关系,最优主题数确定之前需要对数据集中包含的主题数目有一定的先验估计[24]。因此本文结合相关文献的经验法则[4-5,7],估计 LDA 主题数目在 3-8 个,再通过实验分别计算酒店评论与共享住宿评论中 3-8 个主题下的 Coherence 值,利用 Coherence 得分确定最佳的主题数目。确定最佳主题数目之后,本文利用 Python 可视化工具 LDAvis 包对主题下的特征词进行可视化分析。为了确保各个主题之间边界清晰,对于主题词不明确且出现在多个主题的特征词(如"房子"、"房间"等)予以删除,选取频率相对靠前的 8 个词汇作为主题代表,并根据特征词的语义关系归进一步确认主题描述名称。对于特征词与以往文献结果重合度较高的主题,本文结合旅游管理相关文献为主题进行编码归纳,而对于特征词与以往文献结果差异较大的主题,由一组研究人员根据每个主题下特征词列表确定名称,再由另外一组研究人员对该主题的名称进行最后的确认核实。③主题社会网络分析。本部分首先对 LDA 的特征词进行整理归纳,将同一主题下的特征词作为主题特征标识,以此构建主题-主题的外部共现矩阵,见表 1。共词矩阵非对角线上的元素为两两关键词在同一条评论中出现的次数,对角线上元素为该词在所有评论中出现的次数。其次,为了揭示单一主题下特征词关联关系,根据特征词-特征词共现关系构建内部共现矩阵,如表 2。最后,利用 Ucinet 和 Netdraw 软件将携程酒店平台和小猪短租平台的主题社会网络结果进行可视化展示。④主题情感分析。根据 LDA 聚类的结果,对每个主题进行基于 HowNet 词典(8 936)和人工标注(989)方式提取情感词进行情感分析,参照文献[20]的处理方法,将评论文本的情感极性按照积极、中立、消极三元划分。考虑到消费者评论的复杂度和侧重点不同,单条评论可能出现同时评价多个主题的情况,即主题与情感词之间的匹配可能出现一对一、多对一等几种情况。因此,本文参照 W. Duan 等对于酒店评论的处理方法[8],根据标点符号将所有的评论按照单句隔开,匹配[主题特征词,情感词]的句式以便确认各主题的用户情感倾向。以酒店评论中"旁边就是地铁站出行很方便,服务员非常热情,主动问卫生怎么样。"为例,分析过程见图 2。

表 1 主题-主题的共现矩阵示例

主题	1 总体感觉	2 房间硬件	3 设施卫生	4 家庭服务	5 交通便利	6 酒店硬件	7 相处交互
1	12 686	643	2 871	841	2 101	2 122	2 531
2	643	1 984	742	280	428	547	529
3	2 871	742	10 392	919	2 016	1 956	2 587
4	841	280	919	3 520	835	1 306	876
5	2 101	428	2 016	835	7 101	1 669	1 363
6	2 122	547	1 956	1 306	1 669	7 956	1 914
7	2 531	529	2 587	876	1 363	1 914	9 501

注:此表基于携程酒店平台的评论数据

表 2 特征词-特征词共现矩阵示例

特征词	1 价位	2 环境	3 性价比	4 感觉	5 布局	6 总体	7 档次	8 舒适度	9 情调
1	563	84	57	40	2	43	4	1	1
2	84	6 450	332	300	17	142	31	7	15
3	57	332	3 066	205	6	121	24	6	3
4	40	300	205	2 559	23	245	34	4	12
5	2	17	6	23	90	4	1	0	12
6	43	142	121	245	4	1173	8	1	0
7	4	31	24	34	1	8	162	0	0
8	1	7	6	4	0	1	0	50	0
9	1	15	3	12	12	0	0	0	33

注:此表基于携程酒店平台中关于"总体感觉"主题的评论数据

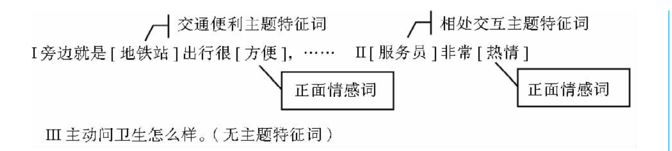


图 2　主题情感分析示例

3　实证结果

3.1　LDA 主题挖掘结果

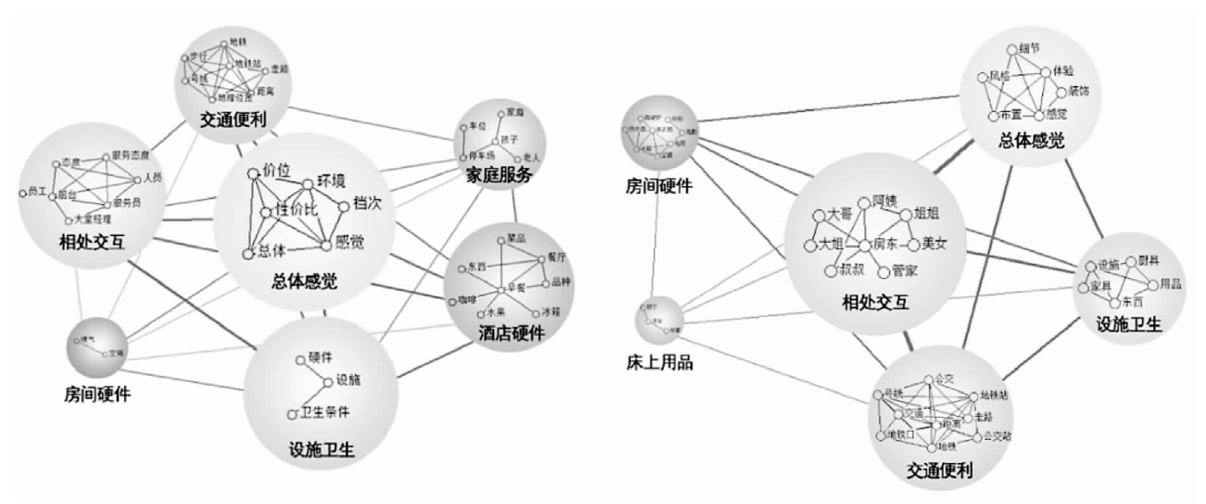
文本的分类主题建立在主题词聚类基础上，根据不同的主题数量可计算出主题一致性程度得分。由于最优主题数的确立需要一定的先验估计，结合旅游管理相关文献[7]，预估酒店预定平台与共享住宿平台的主题数目 3－8 个，本实验进行调试迭代以达到最优聚类结果。由于通过 LDA 模型提取的特征词较多，而过多的主题特征词难以直接用于实际分析[3,25]，因此本研究选取频率靠前的 8 个词汇作为主题代表，再进行主题特征识别归纳。实验结果表明，当携程酒店平台的主题数为 7 时，一致性得分最高(Coherence Score＝0.421)。LDA 模型的结果显示，该平台用户文本评论的七大主题分别为设施卫生、交通便利、房间硬件、相处交互、总体感觉、家庭服务和酒店硬件。当小猪短租平台的主题数为 6 时，一致性得分最高(Coherence Score＝0.420)。LDA 模型的结果显示，该平台文本评论的六大主题分别为设施卫生、交通便利、房间硬件、相处交互、总体感觉和床上用品。其中，设施卫生、交通便利、房间硬件、相处交互和总体感觉五大主题，是携程酒店平台与小猪短租平台中用户共同关注的主题。此外，家庭服务和酒店硬件为携程酒店平台的特色主题，床上用品为小猪短租平台的特色主题。LDA 模型的具体结果如表 3 所示：

表 3　LDA 主题模型结果

序号	主题	携程酒店平台	小猪短租平台
1	设施卫生	设施 酒店设施 环境卫生 卫生条件 硬件 卫生设施 插座	设施 东西 用品 生活用品 物品 家具 厨具 电器
2	交通便利	地铁站 地铁 距离 走路 步行 号线 地理位置	地铁 地铁站 交通 地铁口 公交 距离 走路 公交站 号线
3	房间硬件	空调 浴缸 泡池 电梯 拖鞋 落地窗 窗帘 暖气 电视	电视 冰箱 空调 洗衣机 电影 微波炉 投影 热水器
4	相处交互	前台 服务员 服务态度 大堂经理 人员 员工 态度	房东 姐姐 阿姨 大哥 大姐 美女 叔叔 管家
5	总体感觉	环境 性价比 感觉 总体 价位 档次 布局 舒适度 情调	体验 感觉 风格 格局 装饰 布置 细节
6	家庭服务	停车场 孩子 家庭 小朋友 车位 小孩子 老人	
7	酒店硬件	早餐 餐厅 品种 东西 咖啡 菜品 冰箱 水果	
8	床上用品		床单 被子 毛巾 枕头 被罩 被褥 床垫

3.2　主题社会网络分析结果

本部分进一步基于 LDA 模型的结果构建社会网络，探索两个平台评论文本主题间的关联关系。具体来说，本部分将使用 UCINET 6 软件，并通过共现网络考察主题与主题间的关联关系，以及单一主题下特征词的共现关系。社会网络分析的可视化结果如图 3 所示：



注：左图为携程酒店平台，右图为小猪短租平台

图 3　社会网络可视化图

进一步分析两个平台上用户所关注的有较大重合的主题。①设施卫生主题和交通便利主题下，分别聚焦于总体设施齐全与交通便利程度，两个平台的评论特征词较为相近；②在房间硬件和相处交互主题的用户关注点有略微差异。例如，房间硬件主题下，携程酒店用户聚焦于符合标准化的酒店房间设施提供，比如浴缸、泡池和落地窗等；小猪短租用户则关注于对家庭设施的描述，如冰箱、洗衣机和微波炉等。相处交互主题下，携程酒店用户对于称呼的特征词更为规范统一，如前台、服务员和大堂经理等，而小猪短租用户的特征词为房东、姐姐和管家等，称呼的特征词更为多元；③对于总体感觉主题，携程酒店平台用户更加注重于其价位和档次，小猪短租平台用户更加关注房源风格布置以及入住体验。最后，携程酒店平台与小猪短租平台均具有特色主题，其中①家庭服务和酒店硬件主题专属于携程酒店平台中 LDA 主题模型的结果。家庭服务主题说明在以家庭为单位的旅游出行时，用户更倾向于携程酒店平台，凸显了标准化酒店住宿的优势；酒店硬件主题中的自助餐厅与水果提供服务等也是携程酒店平台用户关注的重点之一。②床上用品主题是小猪短租平台 LDA 主题模型的独有结果，例如床单、被子、枕头、被套等。

首先，图 3 中节点的大小与主题出现的数量成正比，即节点越大，该主题拥有越多的用户关注。从主题与主题的关系来看，携程酒店平台的七个主题中，占比较大的分别是：①以价位、环境和档次等作为特征词的总体感觉主题；②以设施、硬件和卫生条件等作为特征词的设施卫生主题；③以早餐、水果和餐厅等作为特征词的酒店硬件主题。小猪短租平台的六个主题中，占比较大的分别是：①以"大哥、大姐和美女"等作为特征词的相处交互主题；②以公交、地铁和号线等作为特征词的交通便利主题；③以感觉、总体布置等作为特征词的总体感觉主题。

其次，图中主题间连接线条的粗细度与对应节点主题共同出现的数量成正比，可以看出在携程酒店平台的社会网络中，主题与主题间线条粗细程度差异小，表明酒店用户对于平台的关注主题较为平均分散，即酒店预定平台用户的评论内容的关注主题没有明显偏好。而在小猪短租平台的社会网络图中，相处交互、交通便利和总体感觉三个主题的联系尤为密切，表明该平台相关用户对这三个主题尤为关注。

最后，从单一主题下特征词的关系来看，两个平台的交通位置、相处交互和总体感觉三大主题差别不大，而对于其他主题两个平台用户的差异明显。其中，交通便利主题的内部社会网络最为紧密，即公交、地铁站和步行此类特征词同时出现的频次高，表明两个平台用户对交通出行的主题集中在可达性等；在相处交互主题下，小猪短租平台以房东为中心内部的社会网络节点间联系不紧密，而携程酒店平台的节点间连线较多；在总体感觉主题下，携程酒店平台用户关注的是性价比，而小猪短租平台用户关注的是风格与布置；在房间硬件主题下，携程酒店平台用户关注聚焦于空调与暖气的提供，小猪短租平台用户还会关注微波炉、洗衣机、冰箱和投影等；在设施卫生主题下，小猪短租平台用户在携程酒店平台用户关注维度的基础上，还会关注厨具的卫生情况等。在家庭服务主题下，携程酒店平台用户的评论内容主要以孩子为中心，节点连线较少；在酒店硬件主题下，携程酒店平台用户以早餐为中心节点，节点连线较多，说明早餐、菜品与餐厅是该平台用户考虑的重点。在床上用品主题下，是以床单为中心节点，被子和被罩也是小猪短租平台用户同时关注的重要主题。

3.3 主题情感分析结果

主题情感分析将挖掘 LDA 各个主题的情感倾向，进一步识别两个平台上用户评论文本差异。本部分主要通过情感极性去判断评论文本肯定、否定和中立三元情感态度[20]，得出情感得分，见表 4。在得出各个主题情感得分后，依据主题总体占比，利用公式加权平均计算出情感强度，最后按照各平台主题占比排序，相关结果见图 4。

综合各个主题的情感得分，酒店平台评论的积极、消极和中立的情感强度值分别为 0.76、0.06 和 0.18，共享住宿平台情感强度值分别为 0.82、0.05 和 0.11。对比来看，携程酒店平台用户评论中正面情绪得分较低，而小猪短租平台用户正面情绪在各主题间变化波动较大；两个平台在负面情绪总体得分和变化波动几乎无差异。其中，在正面评论中，总体感觉、相处交互和交通便利均为两个平台正面得分较高的主题。而在负面评论中，酒店平台的负面评论主要集中在房间硬件、家庭服务和酒店硬件三个主题；而共享住宿的负面评论主要集中在床上用品、房间硬件和设施卫生上。

表 4　主题情感分析结果

主题	酒店				共享住宿			
	总体占比/%	正面	负面	中立	总体占比/%	正面	负面	中立
总体感觉	26.2	**0.79**	0.04	0.17	20.1	**0.84**	0.05	0.10
相处交互	21.1	**0.83**	0.06	0.11	38.9	**0.86**	0.05	0.09
酒店硬件	15.7	0.70	**0.08**	0.22				
交通便利	14.2	0.77	0.04	0.19	20.9	**0.82**	0.04	0.13
设施卫生	11.6	0.74	0.06	0.20	12.1	0.81	**0.06**	0.13
家庭服务	7.1	0.68	**0.08**	0.24				
房间硬件	4.0	0.54	**0.15**	0.31	5.5	0.67	**0.11**	0.22
床上用品					2.5	0.73	**0.12**	0.15
情感强度		0.76	0.06	0.18		0.82	0.05	0.11

注:情感得分排序前三的加粗表示

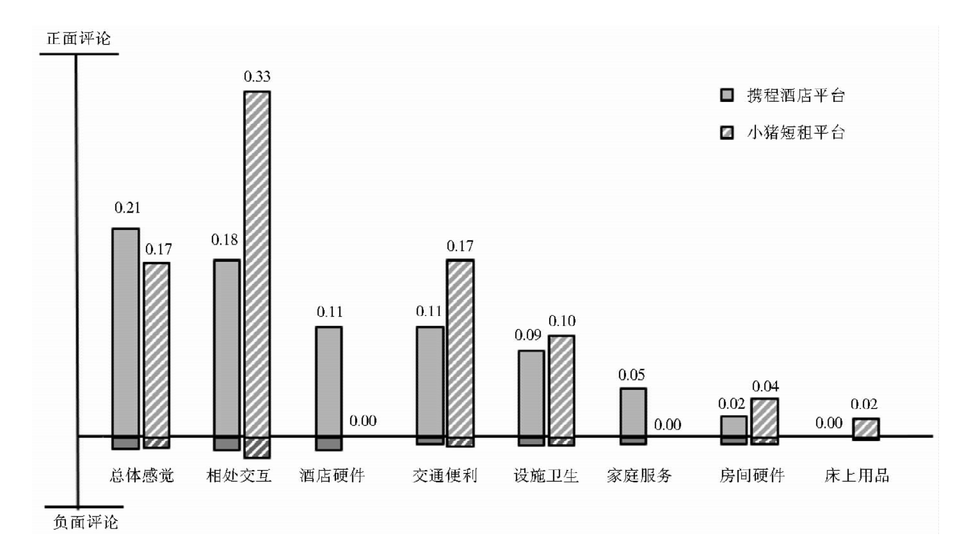


图 4　主题情感强度结果(携程酒店 vs. 小猪短租)

4　结果讨论

4.1　主要发现

基于携程酒店平台和小猪短租平台的用户评论文本数据集，本文通过文本挖掘分析酒店与共享住宿平台上用户评论，首先采用 LDA 模型识别两个平台用户评论的主题;并通过社会网络分析方法将主题词－主题词与主题－特征词的共现网络可视化，进一步分析酒店预定与共享住宿两个平台用户主题间内部网络的差异;最后基于主题情感分析，揭示两个平台上用户情感倾向的具体差异，主要发现如下:①两个平台的文本评论有 5 个共同主题，分别为设施卫生、交通便利、房间硬件、相处交互和总体感觉。共同主题反映的是在住宿体验中平台用户关注的共同内容，例如便利的地理位置和交通对酒店预定平台与共享住宿平台的用户来说都很重要。相关文献也发现了类似的主题内容，例如对于酒店预订平台主题，赵学锋等总结出硬件提供、酒店服务、饮食、性价比和环境五个酒店主题[3]。张梦等基于携程网的实际数据分析，指出酒店设施、服务水平、周边环境对消费者网上预订起到重要作用[26];对于共享住宿平台主题，M. M. Cheng 等基于 Airbnb 文本评论，划分成住宿设施、房东相处和位置等主题[4]。R. Yan 等则基于问卷数据发现对于选择共享住宿的用户而言，服务、便利设施、总体感觉三个因素，均会使游客产生转移意愿[27]。因此，通过跨平台评论共同主题的分析，本文总结酒店与共享住宿平台之间的替代关系体现在设施卫生、交通便利、房间硬件、相处交互和总体感觉等五个共同主题(即用户对于两个平台提供的以上产品或服务存在相似性需求，二者的产品或服务具有替代性)。此外，本文通过对 LDA 聚类的主题挖掘，进一步发现酒店预定平台与共享住宿平台的互补关系(即用户对于两个平台提供产品或服务存在差异性需求，二者的产品或服务具有互补性)。例如在酒店预定平台中，以停车场、孩子、老人等

特征词的家庭服务主题;在共享住宿平台中,以床单、枕头、被子等特征词的床上用品主题。②通过 LDA 的社会网络分析方法深入挖掘两个平台主题词的关联关系,从主题 – 特征词关系可以发现酒店预定平台与共享住宿平台用户即使关注同一主题,但主题下特征词的关联关系仍存在一定的差异。例如相处交互主题下,酒店预定平台用户评论的社会网络节点间联系比共享住宿用户评论更为紧密。以往相关文献针对住宿评论文本处理大多是基于词与词的共现程度,忽视了主题与主题间的关联关系[4,19]。对此本研究引入主题 – 主题的社会网络,进一步分析发现酒店预定平台用户的评论主题内容较为平均分散,类似地,Hou 等基于社会网络分析分析对比携程、途牛与同程的三类酒店平台评论数据,发现同程平台内部话题集中紧密,但主题间较为松散,并指出低关联度网络是由于用户酒店评论关注度不足,对酒店产品服务的接受度较低[28]。与酒店用户相比,共享住宿用户集中关注相处交互、交通便利与总体感觉三个主题。③基于 LDA 主题的情感分析显示两个平台用户评论对各个主题的情绪基本是正面的。其中,总体感觉、相处交互和交通便利均为酒店预定与共享住宿平台正面得分最高的三大主题。两大平台正向情绪得分的主题相似,说明平台用户所需的优质服务和产品基本一致,表明酒店预定平台与共享住宿平台的产品和服务具有一定的替代性;而酒店预定平台的负面评论主要集中在房间硬件、家庭服务和酒店硬件三个主题;共享住宿平台的负面评论主要集中在床上用品、房间硬件和设施卫生上。因此,从主题负向情绪得分的结果发现,两大平台用户所关注的产品或服务主题存在重要差异,表明了酒店预定平台和共享住宿平台所提供的产品和服务也存在一定的互补性。以往文献集中从宏观经济影响的角度讨论共享住宿平台的进入对于当地酒店业绩效的替代或互补作用,且结论有一定争议[29]。对此,本研究则从用户文本评论角度发现了两大平台提供产品和服务同时存在替代性和互补性,对该争议问题进行了解释。

4.2 研究贡献

本研究存在如下三方面的理论贡献。首先,本文通过跨平台的文本评论比较分析,发现了酒店预定平台和共享住宿平台用户评论主题的差异。以往相关文献只针对单一平台,缺乏跨平台的比较分析,仅聚焦于酒店预定平台或共享住宿平台的文本评论进行主题分析[4,7]。其次,本文从微观用户文本评论的角度对比了酒店预定平台和共享住宿平台的上用户评论主题的异同,并探索了用户对两类平台在产品和服务提供上的主题社会网络和情感差异。以往文献主要从宏观经济影响的角度分析共享住宿平台对于传统酒店的影响[1,6],本文进一步从微观用户评论角度分析了两大平台在产品和服务上的替代性和互补性。最后,区别于以往“主题识别 – 共词分析”[19]和“主题识别 – 情感分析”[20]的处理评论文本的方法,本文在主题识别的基础上,构建主题 – 主题和主题 – 特征词的社会网络,主题词 – 情感词匹配来量化用户对于主题的情感倾向的方法流程。该方法流程融合 LDA、SNA 和情感分析方法,细粒度地对两个平台用户评论主题的差异进行了分析。

4.3 实践启示

本文的发现对于相关管理者也有一定的实践启示。首先,通过对携程酒店平台和小猪短租平台上用户文本评论主题的比较分析,总结出两大平台用户共同关注的五大重要主题,即设施卫生、交通便利、房间硬件、相处交互和总体感觉;此外,酒店硬件和家庭服务为酒店预定平台的特色主题,而床上用品为共享住宿平台的特色主题。因此本研究结果能够为相关平台管理者针对其平台用户关心的重要主题进行产品和服务的优化,例如共享住宿的管理者应该重点围绕床上用品进行产品和服务的改进。其次,在主题社会网络分析结果中,酒店预定平台用户的评论内容几乎同时涉及所有主题,共享住宿平台用户的评论内容主要涉及相处交互、总体感觉和交通便利三大主题。不同类型平台的管理者可以用更少的投入获得更显著的回报,例如共享住宿平台房东需要重点突出用户的交互体验,而酒店管理者则需要兼顾产品和服务质量,创造整体良好的用户体验,同时也可以借鉴共享住宿的优势,不拘于标准化为用户提供个性体验,打造充满人情味的酒店。最后,对于旅游住宿平台管理者而言,根据用户期望和出行方式为用户设计量身定做的产品或服务,正视评论中负面信息是旅游住宿保持竞争力的重要途径之一。主题情感分析结果显示酒店平台房间硬件、家庭服务和酒店硬件的情感得分较低;共享住宿床上用品、房间硬件和设施卫生情感得分较低。因此,现阶段两个平台都应将重心放在上述主题所涉及的相关产品和服务,力求以低成本投入换来用户满意度最大提升。具体来说,对于酒店行业而言,相关管理者应该意识到自助餐提供与家庭服务提供的重要性,利用自身优势满足用户需求。对于共享住宿行业,相关管理者应重视房间的清洁度,注意卫生管理,或者可以借用

酒店的标准化管理模式,制定床上用品和设施卫生的评分标准。

5 结语

本文以携程酒店平台小猪短租平台为研究对象,基于 LDA 的主题社会网络和情感分析,探索两大平台上用户文本评论的主题、主题社会网络和主题的用户情感倾向的差异。通过跨平台的用户评论文本分析探索两大主流住宿平台用户平台主题的异同,为酒店预订平台与共享住宿平台的相关管理者有效地进行平台管理提供重要的理论指导和实践借鉴。但是本研究还存在一些不足与需要拓展之处,例如本文没有考虑时间因素的影响,未来可以进一步研究平台用户评论主题的演化机理。

参考文献:

[1] BLAL I, SINGAL M, TEMPLIN J. Airbnb's effect on hotel sales growth[J]. International journal of hospitality management, 2018, 73: 85-92.
[2] XIE K L, KWOK L. The effects of Airbnb's price positioning on hotel performance[J]. International journal of hospitality management, 2017, 67: 174-184.
[3] 赵学锋,汤庆,李岳. 基于客户评论和语料库的在线酒店信誉维度挖掘[J]. 图书情报工作, 2012, 56(12): 124-129.
[4] CHENG M M, JIN X. What do Airbnb users care about? An analysis of online review comments[J]. International journal of hospitality management, 2019, 76: 58-70.
[5] 卢长宝,林嗣杰. 游客选择在线短租住宿的动机研究[J]. 经济管理, 2018, 40(12): 153-167.
[6] DOGRU T, HANKS L, MODY M, et al. The effects of Airbnb on hotel performance: evidence from cities beyond the United States [J]. Tourism management, 2020, 79: 104090.
[7] 吴维芳,高宝俊,杨海霞,等. 评论文本对酒店满意度的影响:基于情感分析的方法[J]. 数据分析与知识发现, 2017, 1(3): 62-71.
[8] DUAN W, YU Y, CAO Q, et al. Exploring the impact of social media on hotel service performance: a sentimental analysis approach[J]. Cornell hospitality quarterly, 2016, 57(3): 282-296.
[9] 李慧,王丽婷. 基于词项热度的微博热点话题发现研究[J]. 情报科学, 2018, 36(4): 45-50.
[10] 夏火松,李保国,杨培. 基于改进 K-means 聚类的在线新闻评论主题抽取[J]. 情报学报, 2016, 35(1): 55-65.
[11] ISHIKAWA S, ARAKAWA Y, TAGASHIRA S, et al. Hot topic detection in local areas using twitter and wikipedia [C]//ARCS 2012. Prague: IEEE, 2012: 1-5.
[12] 唐晓波,邱鑫. 面向主题的高质量评论挖掘模型研究[J]. 现代图书情报技术, 2015(Z1): 104-112.
[13] 涂海丽,唐晓波,谢力. 基于在线评论的用户需求挖掘模型研究[J]. 情报学报, 2015, 34(10): 1088-1097.
[14] 王洪伟,高松,陆颋. 基于 LDA 和 SNA 的在线新闻热点识别研究[J]. 情报学报, 2016, 35(10): 1022-1037.
[15] 叶川,马静. 多媒体微博评论信息的主题发现算法研究[J]. 现代图书情报技术, 2015(11): 51-59.
[16] 黄微,赵江元,闫璐. 网络热点事件话题漂移指数构建与实证研究[J/OL]. 数据分析与知识发现:1-15. [2020-07-18]. http://kns.cnki.net/kcms/detail/10.1478.G2.20200713.1744.012.html.
[17] 冯坤,杨强,常馨怡,等. 基于在线评论和随机占优准则的生鲜电商顾客满意度测评[J/OL]. 中国管理科学:1-12. [2020-07-18]. https://doi.org/10.16381/j.cnki.issn1003-207x.2019.2108.
[18] 叶春蕾,邢燕丽. 基于 LDA 和社会网络中心度的研究生个性化检索推荐模型研究[J]. 图书情报工作, 2015, 59(13): 104-110.
[19] TUSSYADIAH I P, ZACH F. Identifying salient attributes of peer-to-peer accommodation experience[J]. Journal of travel & tourism marketing, 2017, 34(5): 636-652.
[20] 赵常煜,吴亚平,王继民. "一带一路"倡议下的 Twitter 文本主题挖掘和情感分析[J]. 图书情报工作, 2019, 63(19): 119-127
[21] 李煜,刘虹,孙建军. 中国图书馆学博士论文研究主题图谱分析[J]. 图书馆杂志, 2018, 37(6): 22-30.
[22] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in neural information processing systems. Lake Tahoe: Neurips,2013: 3111-3119.
[23] 周建,刘炎宝,刘佳佳. 情感分析研究的知识结构及热点前沿探析[J]. 情报学报, 2020, 39(1): 111-124.
[24] 刘自强,许海云,岳丽欣,等. 基于 Chunk-LDAvis 的核心技术主题识别方法研究[J]. 图书情报工作, 2019, 63(9): 73-84.
[25] 张泰瑞,陈渝. 基于 LDA 模型因素提取的健康信息用户转移行为研究[J]. 图书情报工作, 2019, 63(21): 66-77.
[26] 张梦,张广宇,叶作亮. 在线信息对酒店网上预订的影响研究——基于携程网酒店在线预订数据的分析[J]. 旅游学刊, 2011, 26(7): 79-84.
[27] YAN R, ZHANG K Z K, YU Y. Switching from hotels to peer-to-peer accommodation: an empirical study[J]. Information technology & people, 2019, 32(6):1657-1678.
[28] HOU Z, CUI F, MENG Y, et al. Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis[J]. Tourism management, 2019, 74: 276-289.
[29] ZERVAS G, PROSERPIO D, BYERS J W. The rise of the sharing economy: estimating the impact of Airbnb on the hotel industry [J]. Journal of marketing research, 2017, 54(5): 687-705.

作者贡献说明：
池毛毛：确定选题，提出论文研究框架，论文修改；
潘美钰：数据分析和撰写论文；
王伟军：提出研究思路，修订论文。

A Cross-platform Comparative Study of Reviews on Sharing Accommodation and Hotels Reservation Platform: Combined with LDA-SNA and Sentiment Analysis

Chi Maomao[1,2] Pan Meiyu[1] Wang Weijun[3]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] E-commerce Research Center of Hubei Province, Central China Normal University, Wuhan 430079
[3] Key Laboratory of Adolescent Cyberpsychology and Behavior, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] There may be substitutability and complementarity between shared accommodation and hotel reservation platforms. However, there is still a lack of discussion on which products and services this kind of substitutability and complementarity are embodied in. Therefore, a cross-platform comparative analysis is needed. [**Method/process**] In this paper, 86 635 reviews of relevant rental rooms in Beijing were collected from Ctrip. com and Xiaozhu. com. The cross-platform comparative analysis of online reviews is further carried out by integrated latent dirichlet allocation (LDA), social network analysis (SNA) and sentiment analysis. [**Result/conclusion**] The study found the similarities and differences of the two platforms' users in the comment theme, social network and the emotion of each topic, and further explained the substitutability and complementarity of the two platforms in products and services from the perspective of users' online reviews. The results of this paper also provide important practical reference for platform managers to develop and improve accommodation products and services.

Keywords: cross-platform comparison text topic mining social network analysis sentiment analysis

国家社科基金重大项目“我国政府数据治理与利用能力研究”开题会成功举行

2021 年 1 月 16 日上午，由安小米教授主持的 2020 年度国家社科基金重大项目“我国政府数据治理与利用能力研究”（项目号：20&ZD161）开题报告会以线上会议形式顺利举行。中国人民大学校领导、项目指导专家组成员、项目特邀嘉宾、项目组成员等 50 余人参加了开题会。会议由中国人民大学信息资源管理学院院长刘越男教授主持。项目指导专家组成员有中国科学院计算技术研究所研究员、信息技术战略研究中心常务副主任洪学海，中国人民大学信息资源管理学院院长、教授刘越男，中电长城网际系统应用有限公司教授级高工、ISO、IEC 和 ITU－T 数据安全与隐私保护标准专家闵京华，北京大学政府管理学院教授、北京大学城市治理研究院执行院长沈体雁，南开大学商学院教授、网络社会治理研究中心主任王芳，国家信息中心大数据发展部主任、数字中国研究院院长于施洋，中国人民大学信息资源管理学院教授赵国俊，复旦大学国际关系与公共事务学院教授、数字与移动治理实验室主任郑磊等专家。

中国人民大学副校长顾涛、中国人民大学科研处副处长田洪、《图书情报工作》杂志社副社长杜杏叶、《图书情报知识》副主编（常务）宋恩梅、《电子政务》副主编宋文好、《档案学通讯》编辑部主任张全海、《情报资料工作》编辑石晶作为特邀嘉宾出席。

本次开题会通过搭建政产学研用跨界对话平台，促进了各界关于政府数据治理与利用能力观点、需求和经验的交流，开阔了项目组成员的研究视野，也对进一步完善项目研究计划提供了指导和依据。

（撰稿人：黄婕）